Computational Molecular Biology 2025, Vol.15, No.3, 141-150
http://bioscipublisher.com/index.php/cmb

Research Insight Open Access

Pretrained Language Models for Biological Sequence Understanding

Haimei Wang
Hainan Institute of Biotechnology, Haikou, 570206, Hainan, China
Corresponding author: haimei.wang@hibio.org

Computational Molecular Biology, 2025, Vol.15, No.3 doi: 10.5376/cmb.2025.15.0014
Received: 02 Apr., 2025
Accepted: 13 May, 2025
Published: 04 Jun., 2025

Preferred citation for this article:
Wang H.M., 2025, Pretrained language models for biological sequence understanding, Computational Molecular Biology, 15(3): 141-150 (doi: 10.5376/cmb.2025.15.0014)

Abstract Pre-trained language models (PLMS) are increasingly becoming innovative tools in life sciences, capable of autonomously learning rich representations from massive amounts of biological sequence data. They capture complex patterns and long-term dependencies in DNA, RNA and protein sequences through self-supervised training, effectively compensating for the limitations of traditional bioinformatics methods. This paper reviews the progress of PLM in the field of biological sequence understanding, covering the model principles and their applications in protein function prediction, gene expression regulation, and structural modeling, etc. It focuses on discussing the case of using the ESM-2 model to predict the impact of protein stability mutations and its comparison with traditional methods. Finally, this paper analyzes the challenges such as data sparsity, model interpretability and computational cost, and looks forward to the development prospects of the deep integration of artificial intelligence and molecular biological science. These advancements indicate that pre-trained models are leading a transformation in the research paradigm of biological sequences.

Keywords Pre-trained language model; Biological sequence; Protein function prediction; Gene regulation; Protein structure prediction

1 Introduction

In recent years, with the development of high-throughput sequencing technology, the volume of biological sequence data has grown explosively. The vast repository of DNA, RNA and protein sequences offers unprecedented opportunities for data-driven approaches. However, most of these sequences lack experimental annotations, and the biological laws they contain are difficult to be fully explored through traditional means. The success of natural language processing (NLP) offers an insight into this predicament: if we can "read" biological sequences as we understand human language, it is possible to extract implicit structural and functional information from them. Pre-trained language models were introduced into the field of bioinformatics precisely for this motivation.

They learn the complex relationships between sequence elements through self-supervised pre-training on massive sequences and are regarded as artificial intelligence tools that "understand" the language of life. This method is expected to reveal the "grammatical" rules behind sequences, such as how the combination of amino acids determines protein folding and function, or how the arrangement of nucleotides affects gene regulation (Yang et al., 2023; Wang et al., 2024). In conclusion, the application of language models in biological sequence analysis aims to fully utilize massive data to mine the deep information of biological sequences, providing a brand-new perspective and technical means for understanding life systems.

Traditional bioinformatics methods have many limitations when dealing with biological sequences. First of all, most traditional tools rely on artificially designed features and simplified assumptions. For instance, protein functions are often inferred through homologous sequence alignment, and information on conserved sites is captured using sequence similarity or position-specific scoring matrices (PSSM). However, this strategy based on alignment often fails for sequences lacking known homology and cannot discover functional relationships in distant sequences (Song et al., 2021). Furthermore, some feature representation methods (such as single-hot encoding or fixed-length k-mer fragments) fail to reflect the context association of the sequence, ignoring the remote dependencies and advanced patterns in the sequence. Take the identification of gene regulatory elements

as an example. Traditional methods rely on the scanning of known short sequence motifs, making it difficult to comprehensively consider the impact of a broader sequence background on gene expression. Furthermore, although physical and structural simulation methods (such as protein molecular dynamics or energy functions) are precise, their computational costs are high, making it difficult to apply them on a large scale at the whole-genome or proteome scale. In conclusion, traditional methods are often confined to local or existing knowledge and lack the ability to automatically extract deep patterns from massive sequences, which restricts a comprehensive understanding of the functions of biological sequences. This article aims to comprehensively review the current situation and trends of pre-trained language models empowering biological sequence research, providing useful references for researchers in related fields.

2 Background on Biological Sequences and Representation

2.1 Structure and characteristics of DNA, RNA, and protein sequences

The sequences of biological macromolecules include nucleic acid sequences (DNA and RNA) and protein sequences, which each have their own characteristics in structure and properties. DNA (deoxyribonucleic acid) is composed of four nucleotides (A, T, C, and G), and usually exists in the form of A double helix and double strands. The two opposite parallel strands maintain a stable structure through base pairing (A-T, C-G). DNA sequences carry genetic information and generate RNA through transcription. RNA (ribonucleic acid) is composed of four bases: A, U, C, and G. It is generally a single-stranded structure, but it can form secondary structures such as hairpins locally. RNA in cells not only serves as a messenger for gene expression but also has catalytic or regulatory functions. DNA and RNA sequences mainly function by encoding proteins or regulatory elements (Pan and Shen, 2018).

Protein sequences are composed of 20 kinds of amino acid residues and are the products of gene translation. The type and sequence of amino acids (primary structure) determine how proteins fold into specific spatial conformations (tertiary structure) and thereby perform biological functions. Protein sequences possess diverse chemical properties: The differences in hydrophobicity and charge among various amino acids enable proteins to form complex secondary structures (α -helices, β -folds, etc.) and domain modules. There are often specific conserved motifs or functional domains in sequences that are crucial for protein functions. Therefore, a protein sequence is not merely a string of letters; it also contains rich structural and functional information. Understanding the composition and characteristics of DNA, RNA and protein sequences is the foundation for applying computational models to analyze biological sequences(Helaly et al., 2020).

2.2 Traditional encoding schemes (e.g., one-hot, k-mer, PSSM)

When applying computational models to analyze biological sequences, it is first necessary to convert the sequences into digital representations. Traditionally, researchers have proposed various intuitive coding schemes to represent DNA, RNA or protein sequences.

One-hot encoding is the most fundamental representation method, which represents the identity of each base or amino acid with a high-dimensional sparse vector (for example, a 20-dimensional vector is used for a protein sequence, and only the position of the residue is 1). Single-heat encoding is simple and straightforward, and is often used as the input feature of traditional machine learning models. However, its drawback is that it fails to reflect the similarity between symbols and does not contain any contextual information (Gupta et al., 2024).

The K-mer fragment representation method splits the sequence into consecutive subsequence fragments according to a window of fixed length k, and encodes these fragments as lexical units. For example, DNA sequences can be represented by hexonucleotide fragments of k=6, and each sequence is regarded as a collection of these 6-mer "words". By statistically analyzing the k-mer frequency or mapping the k-mer to an embedding vector, local sequence patterns can be captured to a certain extent. k-mer is widely applied in tasks such as genomic sequence classification and motif discovery. However, it should be noted that k-mer only focuses on local fragments of length k, and long-range relationships beyond the window cannot be reflected (Ng, 2017, Matougui et al., 2020).

The PSSM position-specific scoring matrix uses evolutionary information to represent sequences and is usually constructed through multiple sequence alignment. For a given protein or DNA sequence, first collect the homologous sequences and compare them, count the occurrence frequency of 20 amino acids (or 4 bases) at each position, and form the probability distribution of that position, which is a column of the PSSM matrix. PSSM retains the evolutionary conservation of each position and is more biologically significant than single-heat coding. However, it assumes that each point is statistically independent and it is difficult to express the synergistic changes among different points. In addition, PSSM generation depends on existing homologous sequences, and its effect is limited when dealing with sequences lacking rich homologies (Chia and Lee, 2022).

2.3 Need for contextual representation in biological data

Although the above-mentioned traditional coding methods have promoted sequence analysis to a certain extent, they generally lack the characterization of sequence context dependence. In biological sequences, the influence of a certain base or amino acid on function often depends on its sequence background. For instance, the role of a certain transcription factor binding site in a DNA sequence may be enhanced or weakened by the combination regulation of adjacent sequences. Similarly, whether an amino acid residue constitutes the active site of an enzyme depends on its spatial neighborhood in the tertiary structure of the protein. Static encodings such as single-heat or k-mer cannot assign different representations to the same element according to different environments. This is similar to how in human language, the meaning of a word changes with the context, and simple lexicographical encodings cannot reflect such differences (Fang et al., 2021, Sanabria et al., 2024).

In addition, there are a large number of long-term dependencies and cooperative change patterns in biological sequences. For instance, distant amino acid pairs in proteins can maintain mutual cooperation through coevolution to sustain structural stability or functionality. Traditional feature representations often assume that sequence positions are independent of each other or only consider local fragments, making it difficult to capture such correlation information that spans the entire sequence. This limitation may cause the model to miss key functional clues or make misjudgments (He et al., 2024).

Therefore, it is urgently necessary to introduce methods that can represent the global context information of sequences, so that the representation of each sequence element can dynamically reflect the sequence environment it is in. Such contextualization representations have been proven to be extremely effective in natural language processing and are also highly anticipated in the field of biological sequences.

3 Foundations of Pretrained Language Models (PLMs)

3.1 Overview of NLP-based models: BERT, GPT, and transformers

The rise of pre-trained language models is attributed to the successful application of the Transformer neural network architecture in natural language processing. Transformer achieves efficient modeling of global dependencies in sequences through the self-attention mechanism, where the representation of each position can directly refer to the information of any other position in the sequence. Compared with traditional recurrent neural networks (RNNS), Transformers can process sequences in parallel and capture long-range relationships, thus performing outstandingly in tasks such as language modeling (Kalyan et al., 2021).

Based on this architecture, several landmark NLP models have emerged, among which the representative ones include the BERT and GPT series. BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional encoder model. It is composed of stacked Transformer encoders and pre-trained on large-scale text corpora with a masking language model task, that is, randomly masking some words and then allowing the model to predict missing words, thereby learning the semantic representation of each word in the context. The pre-training of BERT enables it to generate deep context embeddings, and through fine-tuning in downstream tasks such as question answering and classification, it demonstrates an accuracy far exceeding that of previous methods. GPT (generative pre-trained transformer) belongs to the paradigm of autoregressive generative models. The GPT series models (such as GPT-2, GPT-3, etc.) use the Transformer decoder to predict the next word in sequence to train the model, which is a typical language model objective. Due to unsupervised reading of vast amounts of text, the GPT model can naturally generate coherent text and demonstrate astonishing capabilities in

areas such as dialogue generation and writing assistance (Ramprasath et al., 2022). BERT and GPT respectively represent the two major categories of encoder-type and decoder-type in pre-trained models. Their success demonstrates the powerful potential of pre-training combined with the Transformer architecture. Nowadays, the ideas of these two types of models are being borrowed and extended to the field of biological sequences, providing new tools for analyzing DNA, RNA and protein sequences.

3.2 Self-supervised learning tasks adapted for sequences (e.g., masked language modeling, next-token prediction)

The reason why pre-trained models are powerful lies in the fact that they adopt the self-supervised learning strategy, automatically designing training tasks from massive unlabeled data to approximate the essential statistical laws of sequences. When applying the pre-training paradigm of NLP to biological sequences, the training tasks need to be modified accordingly to adapt to the characteristics of DNA, RNA or protein sequences (Figure 1) (Kim et al., 2023).

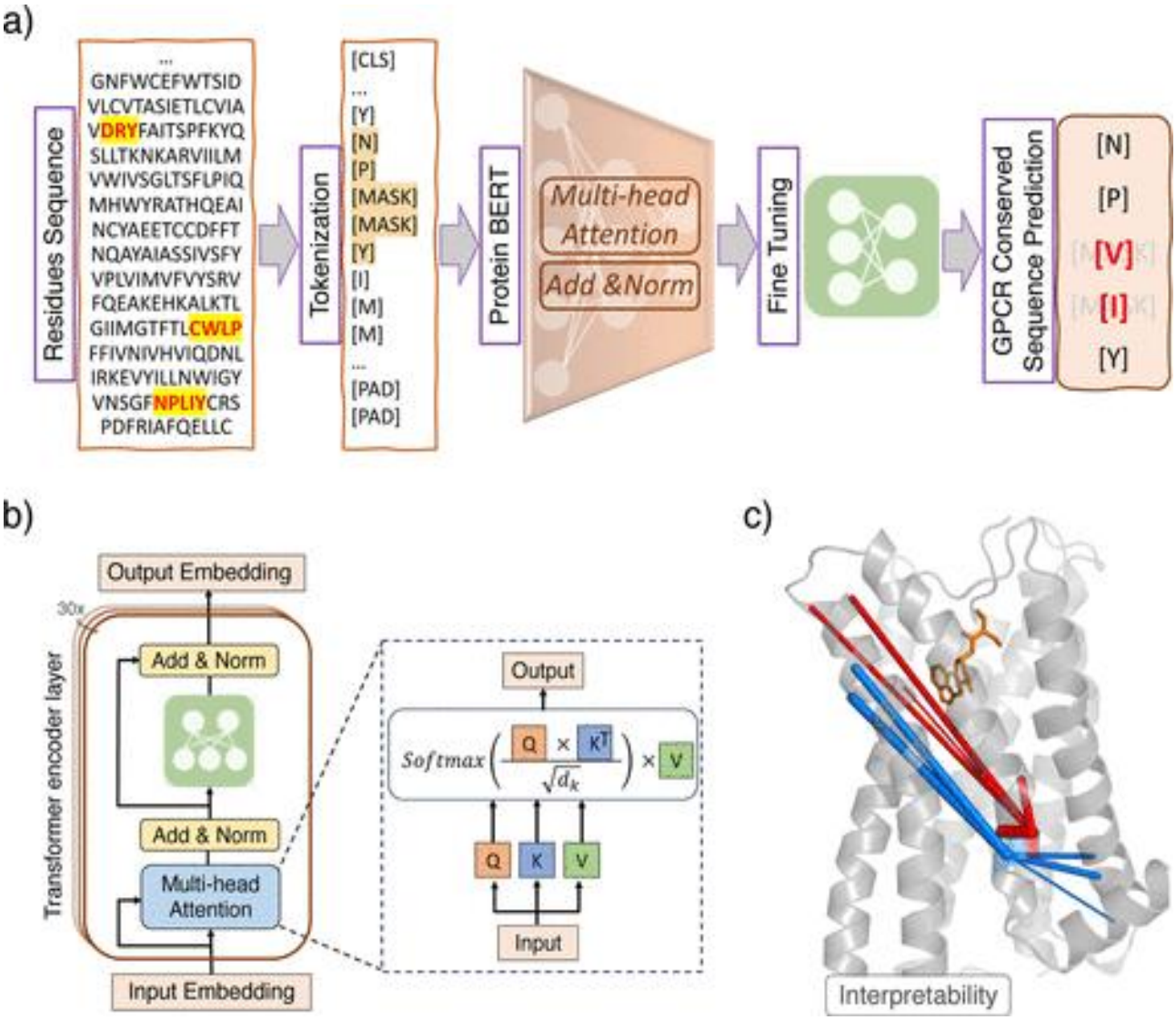

Figure 1 (a) Overall architecture of GPCR-BERT. Amino acid sequences are tokenized and subsequently processed through Prot-Bert, followed by the regression head. (b) Structure of Prot-Bert transformer and the attention layer. The input token embedding is transformed into keys, queries, and values which subsequently form the attention matrix. The output is passed through a fully connected neural network. This sequence of operations is iterated 30 times to reach the final output embedding of the GPCR sequence. (c) Representation of the top five most correlated amino acids to the first x (red) and second x (blue) in NPxxY motif within a GPCR obtained through attention heads. The thickness of the lines represents the strength of correlations (weights) (Adopted from Kim et al., 2023)

Masked language modeling (MLM) is a pre-training task borrowed from BERT. It randomly masks some symbols in the sequence and requires the model to predict the original symbols at the masked positions based on the context(Uribe et al., 2022). In the context of protein sequences, it is common to randomly cover, for example, 15% of amino acids, allowing the model to fill in these gaps, thereby learning the internal constraints and patterns of the sequence. Successfully predicting the masked residue means that the model has captured the restrictive relationship of the remaining sequences to this position (similar to inferential the amino acid at a certain position through other residues in multiple sequence alignment). DNA sequences can also be applied to MLM tasks, such as randomly covering some bases, allowing the model to restore the original bases based on the sequences on both sides, thereby learning the language rules of genomic sequences.

The task of autoregressive sequence modeling (Next-Token prediction) requires the model to gradually predict the Next symbol in the natural sequence order, similar to the training method of GPT. In biological sequences, autoregressive pre-training can be used to construct sequence generation models. For instance, by training a Transformer to predict the next amino acid based on the previous ones, the model gradually learns to generate sequence fragments similar to natural proteins. Such generative models can be used to design entirely new protein sequences or expand partial sequence fragments. For DNA, autoregressive models can simulate the continuation of chromosome sequences, thereby capturing the statistical characteristics of genomic sequences.

These self-supervised tasks make full use of the rich data of biological sequences and can train models without manual annotation. In addition, there are studies exploring other variant tasks, such as masking consecutive fragments, predicting the position of sequence fragments in the entire sequence, or performing denoising and reconstruction on randomly disturbed sequences, etc (Luo et al., 2020). The common goal of these pre-training tasks is to enable the model to learn the internal structure and semantics of sequences to the greatest extent possible without the need for supervised signals, so as to provide a universal representation for downstream biological analysis.

3.3 Differences between natural language and biological sequence modeling

There are some significant differences in modeling between natural language and biological sequences. Firstly, the scale of the basic alphabet of biological sequences is much smaller than that of human language vocabulary (for example, DNA has only four bases). Although the vocabulary can be expanded through the k-mer method, the model needs to capture more combination patterns of a small number of letters. Secondly, biological sequences are often extremely long and lack clear separation structures (such as genomic DNA where thousands of bases are linked together), and models need to handle remote dependencies and implicit hierarchical structures, unlike language sentences which have clear grammatical boundaries. Furthermore, the "semantics" of biological sequences correspond to their biological functions, and the model's understanding of sequences needs to be measured by the performance of downstream tasks. In contrast, the semantic learning of natural language models can be verified by directly judging the meaning or grammatical rationality of sentences. Finally, there is a large amount of homologous redundancy in biological sequence data, which may not only lead the model to lean towards frequent patterns, but also provide an opportunity to utilize evolutionary correlation information (such as simultaneously inputting a group of homologous sequences to learn co-changes). Therefore, while drawing on NLP models, it is necessary to make corresponding adjustments to the model architecture and training strategies in response to these differences to adapt to the unique attributes and challenges of biological sequences.

4 Applications in Biological Sequence Understanding

4.1 Protein function prediction and variant effect analysis

Pre-trained protein language models provide a new approach for protein functional annotation. Traditional methods mainly infer protein functions based on sequence similarity, while large language models can extract deep features from sequences and predict functional attributes even in the absence of obvious homologous sequences. For instance, when the protein sequences generated by the model are embedded for classification tasks, they perform well in distant homology detection and functional domain prediction, significantly enhancing the functional recognition ability of rare proteins (Sun and Shen, 2025).

For point mutations (single amino acid substitutions) on protein sequences, pre-trained models can also be used to evaluate their impact on function. One approach is to calculate the model probability or perplexity changes of the sequence before and after the mutation: if the model considers that the "synchrony" degree of the sequence significantly decreases after the mutation, it suggests that the mutation may disrupt the protein structure or function. This zero-shot scoring has been used to identify potential pathogenic mutations and is consistent with the trend of experimental determination. In addition, fine-tuning training by embedding language models in combination with a small amount of labeled mutation effect data can further improve the prediction accuracy. Overall, pre-trained models offer a new approach for large-scale and rapid mutation risk screening, demonstrating higher sensitivity and applicability compared to methods based on conservative scoring.

4.2 Gene expression and regulatory element prediction

The prediction of gene regulation and expression levels is another important field that benefits from sequence language models. Early methods often relied on pre-identified sequence motifs to search for regulatory elements such as promoters or enhancers, but this approach missed unknown combination patterns. Pre-trained models offer a data-driven approach to automatically learn from genomic sequences which patterns are related to regulatory functions.

Specifically, by using a DNA sequence model that has undergone self-supervised training, we can extract vector representations of fixed-length gene sequences (such as promoter regions or enhancer fragments), and then use them to determine whether the sequence has regulatory activity. Experiments have proved that when such embeddings are used for tasks such as classifying promoters/non-promoters, enhancers/non-enhancers, or predicting splicing sites, their performance is superior to traditional k-mer frequency or motif scanning methods. For instance, the DNABERT model achieved the latest performance at the time in multiple DNA sequence classification tasks, such as identifying elements that promote gene expression, by pre-training masking language models on large-scale genomic sequences. More broadly, genomic language models can also be used to assess the functional impact of mutations in non-coding regions: the difference in the model's scores of sequences before and after mutations can indicate whether the mutation may alter transcription factor binding or chromatin accessibility, thereby affecting gene expression. Although these current methods still need to be verified in combination with downstream experimental data, they have demonstrated great potential in unsupervised exploration of gene regulatory patterns, providing a powerful tool for understanding the "grammar" of non-coding DNA.

4.3 Protein structure and interaction modeling

Pre-trained models can not only handle one-dimensional sequence features, but also demonstrate the ability to infer three-dimensional structures and molecular interactions. Thanks to capturing the implicit evolutionary information in sequences, large-scale language models have become a new approach for rapidly predicting protein structures. Taking Meta's ESM series models as an example, by embedding only a single-sequence Transformer model and combining it with a simplified folding algorithm, ESMFold has been able to directly predict a three-level structure close to experimental accuracy from the sequence. This achievement is remarkable because traditionally, high-precision structural prediction (such as AlphaFold2) relies on co-evolutionary clues provided by multiple sequence alignment, while ESMFold has demonstrated that language models can extract sufficient structural information even without homologous input.

In the modeling of protein-protein or protein-nucleic acid interactions, each sequence can be input into the pre-trained model separately to obtain a representation vector, and then the matching degree of the two vectors (through a classifier or similarity) can be evaluated to determine whether they are likely to bind. Previous studies have utilized this strategy to predict protein-protein interaction partners and achieved good results. Overall, the sequence representations provided by pre-trained models have opened up new directions for the study of molecular recognition and binding. In the future, the combination of them with physical simulation methods is expected to further improve the prediction accuracy of complex interaction interfaces (Esmaeeli et al., 2023).

5 Case Study: Using ESM-2 for Predicting Mutation Impact on Protein Stability

5.1 Overview of ESM-2 and its training on large-scale protein sequences

ESM-2 is a new generation of large protein language model proposed by Meta. Compared with earlier versions, it has a deeper network and more parameters (the maximum version has a parameter number of billions). ESM-2 was unsupervised trained on hundreds of millions of diverse protein sequences and learned rich evolutionary patterns and sequence features. Unlike traditional sequence analysis, ESM-2 can generate high-dimensional representation vectors for any protein sequence without relying on multiple sequence alignment, and performs excellently in downstream tasks such as structure prediction and function prediction. For instance, the ESMFold model utilized the representation of ESM-2 to directly predict the three-dimensional structure of proteins from sequences, demonstrating the pre-trained model's ability to capture the implicit patterns in sequences. This powerful sequence characterization lays the foundation for studying the mutational effects of single sequences (Figure 2) (Pak et al., 2023).

5.2 Application to missense mutations and ΔΔG prediction

The change in protein stability is commonly measured by the ΔΔG value (the difference in free energy caused by mutations, with a positive value indicating a decrease in stability). The effect of single-point amino acid substitution on protein stability can be evaluated by using the ESM-2 model. An effective method is to input the wild-type and mutant sequences respectively into ESM-2 to obtain the embedding vectors, and then train a regression model to output ΔΔG. Because the ESM-2 embedding condenses the multi-level features of the sequence, even when trained on a smaller mutant database, this model can achieve a relatively high accuracy rate. There are also studies that directly compare the difference in probability scores of ESM-2 for sequences before and after mutation as an indicator of stability changes (Zhang et al., 2023). The actual results show that the prediction performance of the model based on ESM-2 on the public mutation stability dataset is comparable to that of specialized supervised learning methods, indicating that the knowledge of the pre-trained model is helpful for accurately capturing the impact of mutations on structural stability.

5.3 Comparison with structure-based and supervised learning baselines

Compared with the physical methods that require a known three-dimensional structure to calculate ΔΔG, the prediction based on ESM-2 does not rely on the experimental structure and has higher computational efficiency. Therefore, it can be extended to proteins with unknown structures and quickly screen for a large number of mutations. Although molecular mechanical energy calculations may be more accurate in the case of providing high-resolution structures, in practice, ESM-2 predictions often achieve comparable accuracy. Compared with traditional supervised learning models, the advantage of ESM-2 lies in the extensive knowledge provided by its pre-training: previous algorithms required manual design of features and training on limited data, with limited generalization ability (Chu et al., 2024). However, the general representations generated by ESM-2 enable reliable results to be achieved even when training on small datasets, significantly improving the robustness of the model. It can be seen from this that the mutation stability prediction driven by ESM-2 has both accuracy and applicability, providing an efficient and reliable tool for studying the effects of protein mutations.

6 Challenges and Limitations

6.1 Data sparsity and imbalance in biological corpora

Although the biological sequence database is large in scale, the distribution of information within it is not uniform. On the one hand, the data volumes of different species and protein families vary greatly, and models tend to focus on learning the dominant patterns in rich data while neglecting rare categories (Ruan et al., 2025). On the other hand, many sequences still lack functional or structural annotations, and the lack of high-quality labeled data makes it difficult for models to be fully trained on certain specific tasks. This kind of data sparsity and imbalance may limit the model's ability to characterize rare functions or novel sequence patterns.

6.2 Model interpretability and biological plausibility

The decision-making process within pre-trained models is often as difficult to understand as a "black box". At present, it is very difficult for us to clearly determine what biological characteristics a certain neuron or attention

weight of the model corresponds to. This lack of interpretability makes it difficult for model predictions to gain full trust from biologists. Furthermore, some of the patterns captured by the model may merely be the correlations of the training data rather than the true causal biological mechanisms, posing a risk of overfitting data bias (Shahid, 2023). Therefore, enhancing the transparency and interpretability of the model's prediction results and verifying their consistency with known biological laws will be a significant challenge in the future.

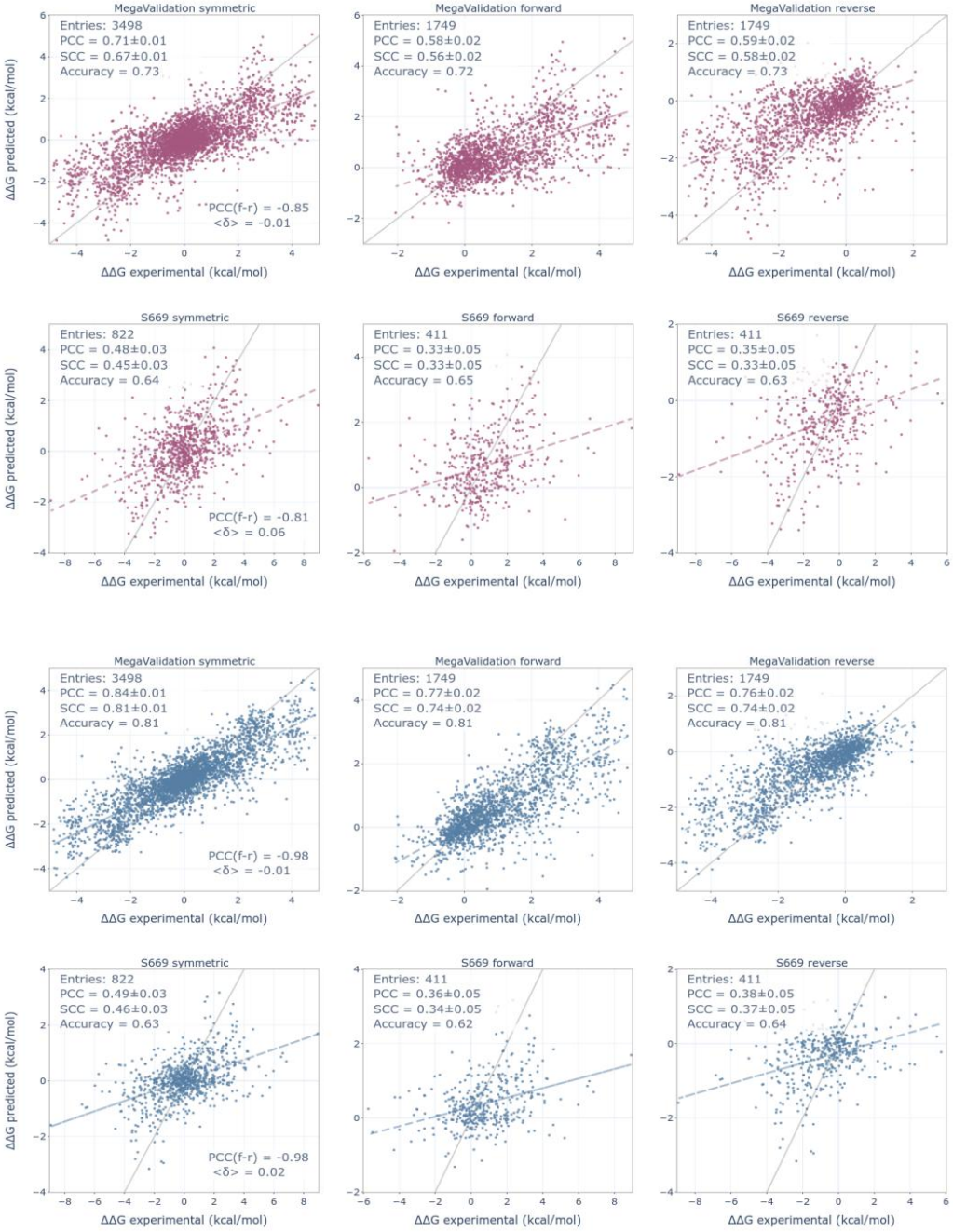

Figure 2 Performance of the model trained on S2648 (pink) dataset and performance of the model trained on MegaTrain (blue) dataset. Same subsets of validation datasets (Adopted from Pak et al., 2023)

6.3 Computational cost and resource constraints

Training and deploying large pre-trained models require huge computing resources and specialized hardware. Training Transformer models with hundreds of millions of sequences not only takes weeks to months but also consumes a large amount of electrical energy. For ordinary research teams, there are resource bottlenecks in

reproducing or fine-tuning such models (Wang et al., 2025). In addition, during the reasoning stage, the prediction of extremely long genomic sequences is also limited by video memory and time. How to improve the computational efficiency of the model and lower the hardware threshold is an urgent problem to be solved for the large-scale application of pre-trained models in bioinformatics.

7 Conclusion

Pre-trained language models have achieved several significant breakthroughs in life sciences. From protein structure prediction to variation effect analysis, such models have demonstrated that deep biological patterns can be decoded merely based on sequence data without the need for artificial feature engineering. They make up for the deficiencies of traditional methods in global context modeling, and have elevated the performance of many biological sequence analysis tasks to a new level.

To consolidate the role of pre-trained models, it is necessary to establish unified evaluation benchmarks and open data. Through objective comparisons on public datasets, the shortcomings of the model can be identified and continuously improved. Meanwhile, sharing large-scale and high-quality sequencing and experimental data will help train more robust models. Continuous benchmark evaluations and dataset construction can ensure that the progress in this field is stable and reliable.

Looking to the future, the integration of artificial intelligence and molecular biology will become increasingly close. Large pre-trained models are expected to become daily tools in biological research: from designing new enzymes and new drugs to real-time monitoring of pathogen evolution, these models will be involved in every aspect of scientific discovery. With the coordinated development of algorithms and experiments, we will step into a new era of molecular biology empowered by AI, revealing the laws of life at a deeper level and accelerating innovation.

Acknowledgments

The author extends sincere thanks to two anonymous peer reviewers for their invaluable feedback on the manuscript.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Chia S.E., and Lee N.K., 2022, Comparisons of DNA sequence representation methods for deep learning modelling, In: 2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), IEEE, pp.1-6.
https://doi.org/10.1109/IICAIET55139.2022.9936754

Chu S.K.S., Narang K., and Siegel J.B., 2024, Protein stability prediction by fine-tuning a protein language model on a mega-scale dataset, PLOS Computational Biology, 20(7): e1012248.
https://doi.org/10.1371/journal.pcbi.1012248

Esmaeeli M., Bauzá A., and Acharya A., 2023, Structural predictions of protein-DNA binding with MELD-DNA, Nucleic Acids Research, 51(4): 1625-1636.
https://doi.org/10.1093/nar/gkad013

Fang G., Zeng F., Li X., and Yao L., 2021, Word2vec based deep learning network for DNA N4-methylcytosine sites identification, Procedia Computer Science, 187: 270-277.
https://doi.org/10.1016/j.procs.2021.04.062

Gupta Y.M., Kirana S.N., and Homchan S., 2024, Representing DNA for machine learning algorithms: a primer, Biochemistry and Molecular Biology Education, 53(2): 142-146.
https://doi.org/10.1002/bmb.21870

He W., Zhou H., Zuo Y., Bai Y., and Guo F., 2024, MuSE: a deep learning model based on multi-feature fusion for super-enhancer prediction, Computational Biology and Chemistry, 113: 108282.
https://doi.org/10.1016/j.compbiolchem.2024.108282

Helaly M., Rady S., and Aref M., 2020, Deep learning for taxonomic classification of biological bacterial sequences, In: Machine learning and big data analytics paradigms: analysis, applications and challenges, Springer International Publishing, pp.393-413.
https://doi.org/10.1007/978-3-030-59338-4_20

Kalyan K.S., Rajasekharan A., and Sangeetha S., 2021, AMMUS: A survey of transformer-based pretrained models in natural language processing, arXiv Preprint, 2021: 103982.
https://doi.org/10.1016/j.jbi.2021.103982

Kim S., Mollaei P., Antony A., Magar R., and Barati Farimani A., 2023, Gpcr-bert: interpreting sequential design of G protein-coupled receptors using protein language models, Journal of Chemical Information and Modeling, 64(4): 1134-1144.
https://doi.org/10.1021/acs.jcim.3c01706

Luo F., Yang P., Li S., Ren X., and Sun X., 2020, CAPT: Contrastive pre-training for learning denoised sequence representations, arXiv Preprint, 2010: 6351.

Matougui B., Belhadef H., and Kitouni I., 2020, An approach based on NLP for DNA sequence encoding using global vectors, In: International Conference of Reliable Information and Communication Technology, Springer International Publishing, pp.577-585.
https://doi.org/10.1007/978-3-030-70713-2_53

Ng P., 2017, dna2vec: Consistent vector representations of variable-length k-mers, arXiv Preprint, 1701: 6279.
https://doi.org/10.48550/arXiv.1701.06279

Pak M.A., Dovidchenko N., Sharma S.M., and Ivankov D., 2023, New mega dataset combined with deep neural network makes a progress in predicting impact of mutation on protein stability, bioRxiv, 31: 522396.
https://doi.org/10.1101/2022.12.31.522396

Pan X., and Shen H., 2018, Learning distributed representations of RNA sequences and its application for predicting RNA-protein binding sites, Neurocomputing, 305: 51-58.
https://doi.org/10.1016/j.neucom.2018.04.036

Ramprasath M., Dhanasekaran K., Karthick T., Velumani R., and Sudhakaran P., 2022, An extensive study on pretrained models for natural language processing based on transformers, In: 2022 International Conference on Electronics and Renewable Systems (ICEARS), IEEE, pp.382-389.
https://doi.org/10.1109/ICEARS53579.2022.9752241

Ruan W., Lyu Y., Zhang J., Cai J., Shu P., Ge Y., Lu Y., Gao S., Wang Y., Wang P., Zhao L., Wang T., Liu Y., Fang L., Liu Z., Li Y., Wu Z., Chen J., Jiang H., Pan Y., Yang Z., Chen J., Liang S., Zhang W., Ma T., Dou Y., Zhang J., Gong X., Gan Q., Zou Y.X., Chen Z.C., Qian Y., Yu S.R., Lu J., Song K., Wang X., Sikora A., Li G., Li X., Wang Y., Zhang L., Abate Y., He L., Zhong W., Liu R., Huang C., Liu W., Shen Y., Ma P., Zhu H., Yan Y., Zhu D., and Liu T., 2025, Large language models for bioinformatics, Quantitative Biology, 14(1): e70014.
https://doi.org/10.1002/qub2.70014

Sanabria M., Hirsch J., Joubert P.M., and Poetsch A., 2024, DNA language model GROVER learns sequence context in the human genome, Nature Machine Intelligence, 6(8): 911-923.
https://doi.org/10.1038/s42256-024-00872-0

Shahid U., 2023, Leveraging fine-tuned large language models in bioinformatics: a research perspective, Qeios, 10: 32388.
https://doi.org/10.32388/WE7UMN.2

Song B., Li Z., Lin X., Wang J., Wang T., and Fu X.Z., 2021, Pretraining model for biological sequence data, Briefings in Functional Genomics, 20(3): 181-195.
https://doi.org/10.1093/bfgp/elab025

Sun H., and Shen B., 2025, Structure-informed protein language models are robust to missense variants, Human Genetics, 144(2): 209-225.
https://doi.org/10.21203/rs.3.rs-3219092/v1

Uribe D., Cuan E., and Urquizo E., 2022, Fine-tuning of BERT models for sequence classification, In: 2022 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE), IEEE, pp.140-144.
https://doi.org/10.1109/ICMEAE58636.2022.00031

Wang N., Bian J., Li Y., Li X., Mumtaz S., Kong L., and Xiong H., 2024, Multi-purpose RNA language modelling with motif-aware pretraining and type-guided fine-tuning, Nature Machine Intelligence, 6(5): 548-557.
https://doi.org/10.1038/s42256-024-00836-4

Wang Z., Wang Z., Jiang J., Chen P., Shi X., and Li Y., 2025, Large language models in bioinformatics: a survey, arXiv Preprint, 2025(v1): 4490.
https://doi.org/10.18653/v1/2025.findings-acl.184

Yang W., Liu C., and Li Z., 2023, Lightweight fine-tuning a pretrained protein language model for protein secondary structure prediction, bioRxiv, 22: 530066.
https://doi.org/10.1101/2023.03.22.530066

Zhang Y., Gao Z., Tan C., and Li S.Z., 2023, Efficiently predicting protein stability changes upon single-point mutation with large language models, arXiv Preprint, 2312: 4019.
https://doi.org/10.48550/arXiv.2312.04019